Journal of Environmental Science, Health & Sustainability, 2025, 1(2), 144–159
https://journals.enviromindsolutions.com/jeshs
eISSN: 3067-946X

Article

Analyzing the effects of data variability and quantity on predicting particulate matter ($PM_{2.5}$) concentrations: Insights from a machine learning approach

Jada A. Macharie[1,*], Wenge Ni-Meister[1] and Maddalena Romano[1]

[1]Department of Geography and Environmental Science, Hunter College of the City University of New York, New York, NY 10021, USA

How to cite

Macharie, J.A., Ni-Meister, W., Romano, M., 2025. Analyzing the effects of data variability and quantity on predicting particulate matter ($PM_{2.5}$) concentrations: Insights from a machine learning approach. Journal of Environmental Science, Health & Sustainability, 1(2), 144–159. https://doi.org/10.63697/jeshs.2025.10042

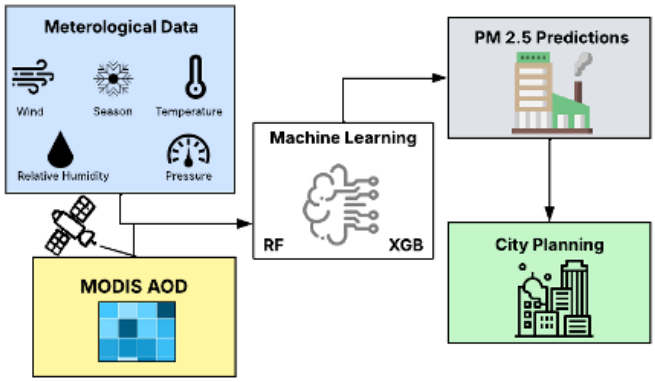
Article info

Received: 17 July 2025
Revised: 25 August 2025
Accepted: 28 August 2025
Published: 30 August 2025

Keywords

$PM_{2.5}$ concentrations
MODIS AOD
Air quality management
Artificial intelligence
Predictive modelling

Graphical abstract



Highlights

- *Aerosol optical depth (AOD) values within a 10 km radius of monitoring stations were most effective in addressing cloud cover interference.*
- *AOD values in conjunction with meteorological data were suitable for $PM_{2.5}$ predictions.*
- *The integration of satellite and ground-based data was essential for reliable $PM_{2.5}$ estimation.*
- *The developed machine learning models demonstrated generalizability in estimating $PM_{2.5}$ levels at other locations.*

Abstract

Accurately predicting particulate matter, 2.5 microns or less in diameter ($PM_{2.5}$), concentrations is imperative to the future of public health and environmental policies. Machine learning models incorporating spatial and temporal datasets to predict $PM_{2.5}$ concentrations are often limited by data availability and poor-resolution satellite imagery. In this study, we present multiple predictive models designed for generalized $PM_{2.5}$ predictions, the output of which has been utilized for different spatial locations. Using Random Forest (RF) and Extreme Gradient Boost (XGB) algorithms, these predictive models follow a multidisciplinary approach using Moderate Resolution Imaging Spectroradiometer Aerosol optical depth (MODIS AOD) and surface datasets (relative humidity, barometric pressure, outdoor temperature, wind speed and wind direction). Models are trained and validated based on historical data to evaluate the impact of training data variability and quantity on the predictive performance of RF and XGB models for $PM_{2.5}$ concentrations. Using MODIS AOD alone yielded weak predictive performance, with average R^2 values ranging from -0.06 to 0.07 across the three urban areas (Washington, D.C., Boston, and New York City), highlighting its limited capability. The integration of meteorological data (temperature, wind speed, wind direction, relative humidity, and barometric pressure) along with MODIS AOD significantly improved the model performance. RF

*Corresponding author: JADA.MACHARIE96@myhunter.cuny.edu (JAM)
© 2025 The Authors. Published by Enviro Mind Solutions.
Handling Editor: Dr. Mahmudur Rahman with assistance from Dr. Sharon Kahara

models achieved R^2 values of 0.30–0.62, while XGB models had R^2 values of 0.25–0.63, with corresponding RMSE values reduced by 20–30% relative to AOD-only models. Feature importance analysis revealed that $PM_{2.5}$ predictions were most strongly influenced by temperature (average importance of 0.21), wind speed (0.20), and wind direction (0.15). MODIS AOD exhibited moderate importance (≈0.12), indicating that although satellite-based aerosol observations contributed to the predictions, ground-based meteorological variables remained the primary drivers. These quantitative results highlighted that combining satellite observations with meteorological measurements substantially enhanced $PM_{2.5}$ predictive accuracy, informing urban planning, environmental policy, and public health interventions to better protect vulnerable populations.

1 Introduction

Particulate matter, 2.5 microns or less in diameter ($PM_{2.5}$), commonly sourced from motor vehicles, burning of fossil fuels, and power plants-have a wide range of detrimental effects on human health (American Lung Association, 2024). Studies show that traffic-related air pollution alone contributes significantly to urban $PM_{2.5}$ concentrations (Qin et al., 2006; Karner et al., 2010). Exposure to these fine particles can cause both short-term and long-term health effects such as, coughing, sneezing, shortness of breath, bronchitis, and long-term health effects like lung cancer, chronic obstructive pulmonary disease (COPD), stroke, and ischemic heart disease (American Lung Association, 2024; Feng et al., 2016). Research has shown that people of color are disproportionately affected by poor air quality. Tessum et al. (2019) found that non-Hispanic White individuals generate approximately 17% more $PM_{2.5}$ than they consume, while Black and Hispanic groups inhale 56% and 63% more, respectively, then they are responsible for producing.

A number of studies have attempted to predict $PM_{2.5}$ concentrations using atmospheric data, most commonly Moderate Resolution Imaging Spectroradiometer Aerosol Optical Depth (MODIS AOD). Kumar et al. (2007) observed a significant positive association between AOD and $PM_{2.5}$ concentrations at both point level (disaggregated) and 5–10 km AOD pixel levels (aggregated). Although AOD has shown strong predictive capabilities due to its positive correlation with $PM_{2.5}$ concentrations, it cannot always be relied upon as the sole predictor in models (Chu et al., 2016). Relying solely on AOD poorly predicts $PM_{2.5}$ concentrations due to the inconsistency of satellite imagery availability over a longer period. Some studies overcome this issue by integrating more variables into their predictive models. Kibirige et al. (2023) aimed to accurately predict $PM_{2.5}$ concentrations in Northern Taiwan using data from air quality monitoring stations. In addition to traditional monitoring data, they incorporated remote transported pollutants (RTP) variables to capture the influence of air pollutants transported from other regions. Their neural network models were trained on data from 2014–2015 and tested on 2016 data. The evaluation included two different datasets: the Extended Local Satellite Dataset (ESD), which provided daily-level $PM_{2.5}$ predictions, and the RTP-based dataset, which offered hourly-level $PM_{2.5}$ predictions (Kibirige et al., 2023). High accuracy within models were attained when they were fed with meteorological data only (Kibirige et al., 2023). A modeling approach using Extreme Gradient Boost and Inverse-Distance Weighting with MODIS AOD, meteorological data, and land use showed strong performance in estimating mean and maximum $PM_{2.5}$ concentrations, with mean absolute errors of 3.68 and 9.20 $\mu g/m^3$ and mean absolute deviations from the median of 8.55 and 15.64 $\mu g/m^3$ (Gutiérrez-Avila et al., 2022).

Several studies have already effectively applied machine learning algorithms/models to predict ambient $PM_{2.5}$ concentrations at high spatial resolutions using satellite-derived AOD values (Paciorek and Liu, 2009; Kumar and Pande, 2023). Currently, the use of machine learning models in air quality modeling had expanded significantly, integrating various approaches such as convolutional neural network (CNN), random forest (RF), Extreme Gradient Boost (XGB), and deep learning (DL) models. Many models face issues with overfitting, temporal bias, and limited generalizability across cities. A pipeline combining a CNN and RF model with local contrast normalization was developed to detect $PM_{2.5}$ hotspots at 300 m resolution using satellite imagery and meteorological data. The CNN extracted predictive features from the imagery, which were then used in the RF model along with meteorological inputs to generate the final predictions (Zheng et al., 2021).

Random forest and XGB provide many benefits in air quality modeling. Random forest generates an ensemble of decision trees and aggregates their outputs to produce a final prediction (Nath et al., 2022). Sample features were termed column sampling and data points as row sampling (Samad et al., 2023). While XGB iteratively improves predictions by minimizing errors across the model demonstrating superior optimization capabilities for pollutant concentration prediction (Li et al., 2022). Scientists estimated daily CO concentrations in Taiwan for the period from 2000 to 2018 using deep neural network, RF, and XGB. They concluded that XGB had the highest R-squared (R^2) values of 0.85, followed by RF and neural network with 0.84 and 0.81, respectively (Wong et al., 2021). Additionally, day of the week and season were also considered in studies predicting $PM_{2.5}$ concentrations to increase model performance. Day of the week and season were highlighted as the most critical predictors, because of their strong connection with weather-linked seasonal effects (Kaveh et al., 2025). When combined with meteorological and AOD data, these predictors were consistently identified as the most important across diverse modeling approaches, including RF, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM).

The goal of this study was to build on existing knowledge by applying RF and XGB models to generate a generalized machine learning outcome for comparing $PM_{2.5}$ predictions across three cities (McMillian, Washington, D.C.; Queens, New York City; and Dudley Roxbury, Boston, Massachusetts) with varying levels of urban characteristics along the East Coast of the United States of America (USA). This study also assessed how increased data variability and quantity affected machine learning models in accurately predicting air quality by integrating remote sensing data and ground measurements. Furthermore, beyond the prediction of $PM_{2.5}$ concentrations, this study also explored the mechanisms driving air quality dynamics and generalizability of models to be utilized for other cities.

2 Study area

The study area is located in three major cities of the USA, which includes air monitoring stations in McMillan, Washington, D.C., Queens, New York City, and Dudley Roxbury, Boston, Massachusetts (**Fig. 1**). All three locations were among the many areas in the United States (US) that have taken initiatives toward cleaner air, by partnering with the Department of Energy and Environment (DOEE), USA.

Both $PM_{2.5}$ and meteorological sensors were located above ground to collect data that was representative of the area, negating ground influences such as road dust and human activity. The height of the sensors also supported comparisons across locations, such as urban vs. rural or high density vs. low density areas. $PM_{2.5}$ sensors were located 3 to 4 m above ground level (agl), temperature sensors were located 1.5 to 2 m agl, relative humidity sensors were located 1.5 to 2 m agl, wind speed and direction were located 10 m agl, and barometric pressure was located at the station elevation level. Wind speed and wind direction sensors were relatively higher than the rest to reduce interference from buildings and trees.

3 Methodology

3.1 Datasets

Meteorological variables, $PM_{2.5}$ measurements and MODIS AOD were collected for this study. AOD values were retrieved from MODIS Terra/Aqua MCD16A2 satellite, utilizing Google Earth engine. Daily values were taken within the 10 km buffer around each monitoring station (McMillian, Queens, and Dudley Roxbury) for the temporal period (2011–2023). A 10 km buffer was applied for MODIS AOD data collection to align with the sensor's spatial resolution, which estimates aerosol concentrations over a 10 km × 10 km area (Remer et al., 2005).

Meteorological variables collected using United States Environmental Protection Agency's (US EPA's) Air Quality System (AQS) database, include temperature (T, °F), relative humidity (RH, %), barometric pressure (PP, mbar), wind speed (WS, knots) and wind direction (WD, degrees). For each variable, the maximum (max) value and corresponding hour were also recorded. These datasets were used as inputs for model simulations employing RF and XGB to generate predictions (**Table 1**).

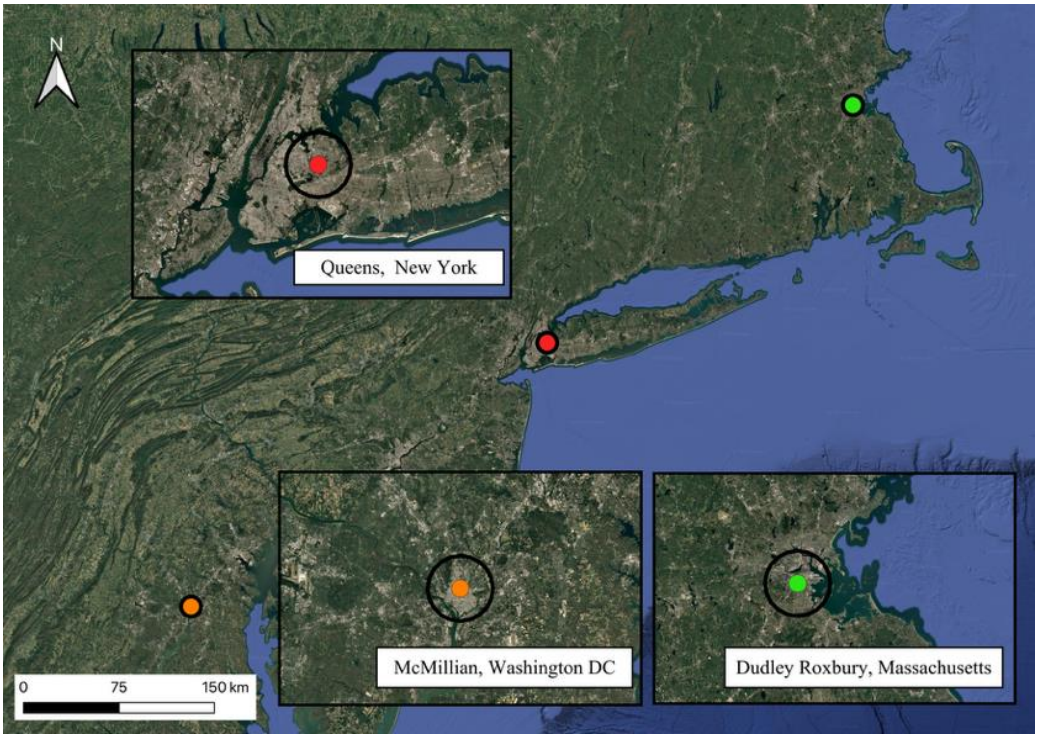


Figure 1. *Map of the study area in three cities in the Northeastern USA. Circles represent a 10 km buffer.*

Table 1. *Datasets used for model developments.*

Variable	Description	Source	Units
$PM_{2.5}$	Ground-level particulate matter concentration	US EPA AQS database	μg/m³
AOD	Aerosol optical depth	MODIS Terra/Aqua MCD16A2 via Google Earth Engine, daily mean within 10 km buffer	Unitless
T	Temperature (Max, mean values and max hour)	US EPA AQS database	°F
RH	Relative humidity (Max, mean values and max hour)	US EPA AQS database	%
PP	Barometric pressure (Max, mean values and max hour)	US EPA AQS database	mbar
WS	Wind speed (Max, mean values and max hour)	US EPA AQS database	knots
WD	Wind direction (Max, mean values and max hour)	US EPA AQS database	degrees

3.2 Modelling technique

Two Artificial Intelligence (AI) models were used to identify the ideal model for predicting the concentrations of $PM_{2.5}$, based on meteorological conditions. Random forest model was built on decision tree structure by combining the outputs of multiple decision trees to enhance predictive accuracy. Each decision tree in the forest was trained on a random set of data, using the bagging technique. This technique reduces the risk of overfitting that can with individual decision trees, as it averages out the noise and variability in the data. Random forest introduces randomness that lacks collinearity, making the overall model generalized and capable of handling complex patterns in the data (Park et al., 2020). To predict $PM_{2.5}$ concentrations, RF model effectively manages the non-linear relationships between $PM_{2.5}$ concentrations and environmental variables and AOD. On the other hand, XGB model uses an advanced form of boosting to build models sequentially, focusing on reducing errors made from previous iterations. Each model prioritizes correcting the remaining errors of the earlier models by assigning greater weight to poorly predicted data points. The final output was a weighted combination of all the models. Like RF, XGB model effectively captures complex pattern in the data while minimizing the errors (Chen and Guestrin, 2016).

Random forest and XGB models have many similarities and differences. Both effectively capture data patterns and relationships in complex scenarios based on their assigned parameters. This study does not utilize hyperparameters tuning within RF and XGB model structures. Lack of hyperparameters allows for generalizations and future model integrations.

3.3 Modelling framework

Figure 2 outlines the detailed modelling framework utilized in this study. Data preparation, including data gathering, and data cleaning was a significant step of modeling efforts. Data preparation includes data organization and data cleaning steps, such as deleting missing data to produce a dataset without NaN, not a number, values. This step ensures combining AOD values, meteorological data and $PM_{2.5}$ concentrations with the same corresponding dates and times.

Data preparation and cleaning were essential steps in ensuring the accuracy and reliability of the $PM_{2.5}$ prediction models. The process began with collecting data from multiple sources (**Table 1**). Ground-based $PM_{2.5}$ measurements were obtained from air quality monitoring stations, providing direct observations of fine particulate matter concentrations. MODIS AOD values were retrieved from the Terra and Aqua MCD16A2 satellites using Google Earth Engine, with daily values extracted within a 10 km buffer surrounding each monitoring station. This buffer aligns with the satellite sensor's spatial resolution, which estimates aerosol concentrations over a 10 km × 10 km area. Meteorological variables, including temperature, relative humidity, barometric pressure, wind speed, and wind direction were acquired from the US EPA Air Quality System (AQS) database, providing hourly measurements at each station.

A key challenge in data preparation was reconciling datasets with different temporal resolutions. While AOD data were available at a daily temporal scale, meteorological and $PM_{2.5}$ measurements were recorded hourly. To address this, hourly $PM_{2.5}$ and meteorological data were aggregated into daily averages to match the AOD data. This ensured

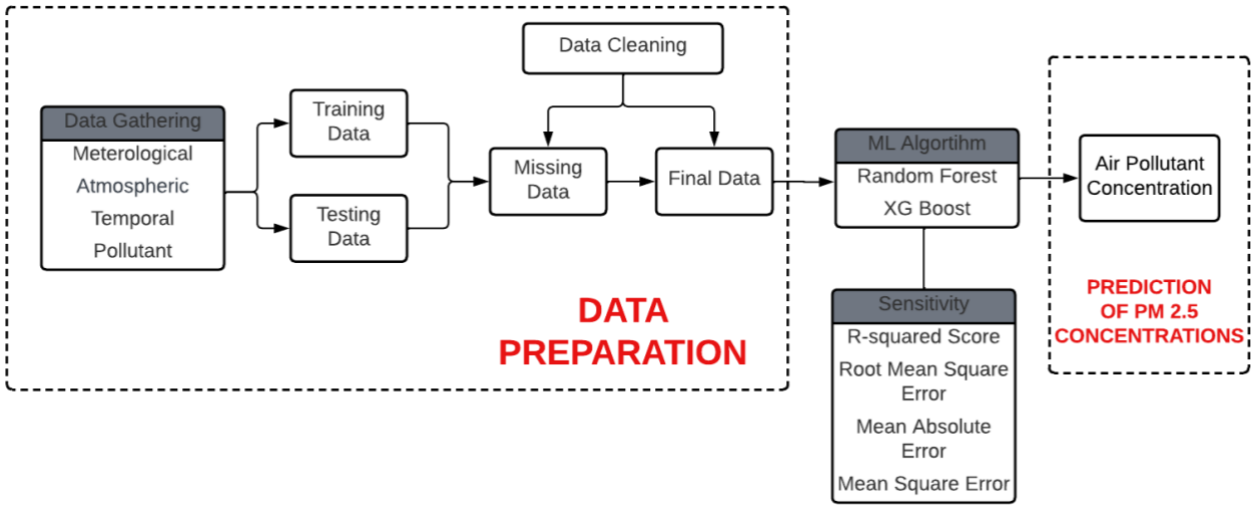


Figure 2. *Methodological framework of atmospheric and surface datasets and steps taken to predict $PM_{2.5}$ concentrations.*

that each observation in the combined dataset corresponded to a consistent daily temporal unit, allowing the models to learn relationships between daily AOD, meteorological variables, and $PM_{2.5}$ concentrations effectively.

Data cleaning was conducted to handle missing or invalid values, which can arise from instrument malfunctions, cloud cover affecting satellite observations, or gaps in station records. All missing values, represented as NaN, were identified across all variables. Rows containing incomplete observations were removed to produce a complete dataset without missing values. Although imputation methods could have been applied, removing rows with missing data was preferred to avoid introducing potential biases, particularly given the sensitivity of machine learning models to inconsistent inputs.

Once cleaned, aligned, and aggregated to the daily scale, the datasets were merged into a single structured dataset suitable for input into RF and XGB models. Each row of the final dataset contained the daily average $PM_{2.5}$ concentration, the corresponding daily AOD value, and aggregated meteorological variables for that day and location. This comprehensive and consistent dataset provided a strong foundation for the machine learning models to capture complex interactions between AOD, meteorological conditions, and ground-level $PM_{2.5}$ concentrations, ultimately improving prediction accuracy.

Python version 3.10.9 and Google Earth Engine (GEE) was employed to import data, spatially clean data, perform statistical analysis, and build and run models. A baseline model was first established to evaluate performance improvements through iterations prior to prediction simulations. To ensure model accuracy, the framework integrates a performance analysis, using metrics such as R^2 values, Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE). For the model development, the input data included the following parameters: AOD, T, Tmax, RH, PP, WS, WD, day, month, and year.

For the modeling, 80% of the available data was used for training, while the remaining 20% was utilized for testing. This split ensured that the training phase captured a broad range of conditions, while the test phase evaluated model performance on unseen data. This methodology was consistently applied across different study area locations, with models built, trained, and tested using data from Washington, D.C., Boston, and New York City. $PM_{2.5}$ predictions were generated using models trained on either five or ten years of data, incorporating MODIS AOD and meteorological variables for each city. Additionally, the same model structure was applied to predict $PM_{2.5}$ concentrations for one city using the meteorological and AOD data from the other cities, for example using Washington, D.C. and Boston data to predict New York City $PM_{2.5}$ concentrations.

3.4 Model evaluation metrics

Feature importance functions were utilized to understand which predictor variables play a major role in $PM_{2.5}$ prediction and variability. In RF and XGB models, feature importance was calculated by measuring the contribution of each feature to reducing variance across all the decision trees in the group. Features that were consistently led to greater reductions in predictions error were given higher importance scores, thereby highlighting their influence on the models' output.

A performance evaluation of these models was very crucial in understanding the key contributors to the models' performances. Weighing more on variables that contribute positively to the models can drastically increase model accuracy. Additionally, being able to see how outputs vary

with inputs, exposes the behavior and robustness of the models.

4 Results

4.1 MODIS AOD

For visualizing long-term trends in AOD, daily MODIS AOD values were smoothed using a 30-day moving average to reduce short-term variability and improve readability (**Fig. 3**). This moving average calculates the mean AOD over a 30-day window that shifts sequentially across the dataset, highlighting broader temporal patterns while minimizing the impact of daily fluctuations.

From 2011 to 2021, daily MODIS AOD observations were extracted within a 10 km buffer around each monitoring site. Over the 10-year period, a total of 3,227 daily AOD observations were recorded for Washington, D.C.; 2,511 for Boston; and 2,532 for New York City. Each value represented a single daily observation at each site, and the dataset included only days with available MODIS measurements. Washington, D.C. had the maximum value of 0.3815 on 07/26/2021 and a minimum value of 0.0598 on 12/08/2015. In Boston, a maximum value of 0.3052 on 08/23/2021 and a minimum value of 0.0872 on 12/19/2015 was recorded, while a maximum value of 0.3531 on 08/11/2021 and a minimum value of 0.0965 on 02/10/2012 were recorded in New York City (**Fig. 3**). Washington, D.C. had the highest recorded AOD value among three monitoring stations, indicating more atmospheric particulate matter than New York City and Boston. Washington, D.C.'s peak AOD value in 2021 suggests a potential increase in pollution events such as extreme weather (Liu et al., 2022). All three cities had minimum AOD values below 0.1 in the late 2010s, indicating consistent periods of low particulate matter in the atmosphere. The AOD peaked during the late 20th century in all three cities could reflect historical industrial and vehicular emissions, AOD peaks may specifically be attributed to the Canadian wildfires in 2021 (Jaffe et al., 2020).

Typically, AOD values were often elevated in the winter, variable in the spring, highest in the summer and lowest in the fall. AOD levels were the highest in the Summer, especially in urban areas because biomass/fossil fuel burning increases with the uses of cooling equipment and vehicles, which increases aerosol loading. Whereas the cooling temperatures in the fall reduces aerosol production. Colder temperatures in the winter cause elevated AOD levels because of the use of residential heating, which increases emissions and consequently raises aerosol loading (Gupta and Christopher, 2008). All three cities have recorded similar ranges of AOD values, frequently fluctuating between 0.1 and 0.5. This shows a baseline level of aerosol presence across the urban environments. New York City stands out as the only city that has experienced higher peaks of AOD levels over the 10 years (**Fig. 3**).

In this study, AOD measurements were recorded at 550 nm due to their sensitivity to $PM_{2.5}$. This wavelength, located in the green portion of the visible spectrum, effectively captures the scattering and absorption properties of aerosols, with minimal interference from atmospheric gases. As a result, it provided a reliable substitute for assessing $PM_{2.5}$ concentrations in urban environments.

4.2 Environmental variables

Over the 10-year period, extreme meteorological values did not consistently group within a specific decade or cluster of years, suggesting temporal variability in climatic extremes. However, during few years, specifically 2011, 2013, 2015, 2016, 2017 and 2019, frequent extremes were observed,

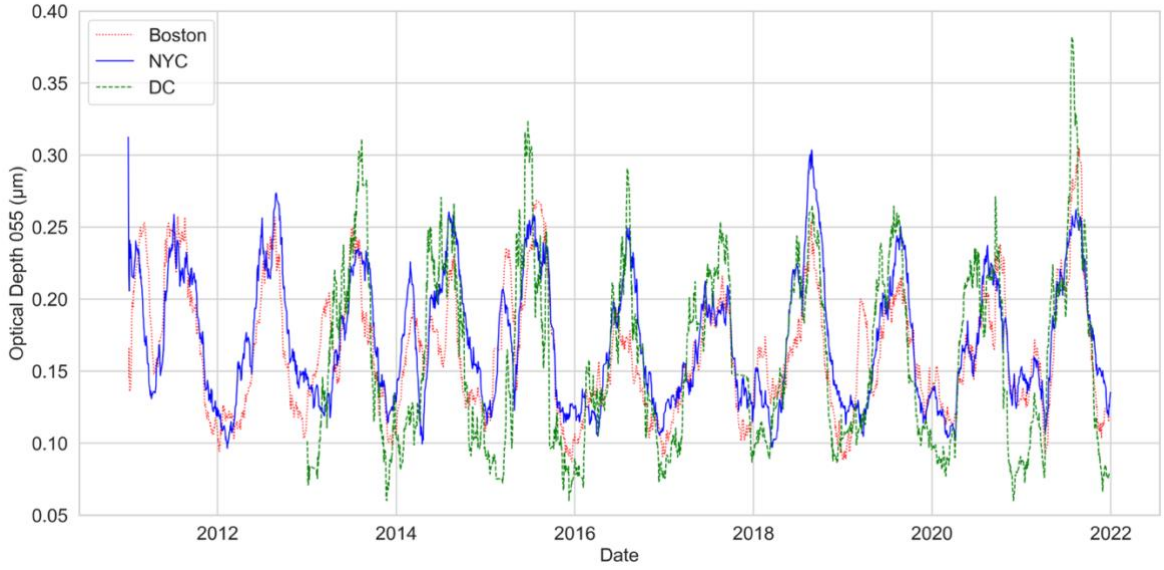


Figure 3. *Time series of MODIS AOD for 2011–2021. Data was smoothed using a 30-day moving average to make increase readability.*

indicating increased atmospheric variability during these periods. While some cities showed seasonal patterns, such as higher wind speeds in summer for New York City and Washington, D.C., extreme values overall vary by city and weather variable, highlighting the region's complex and dynamic climate behavior (**Table 2**).

Furthermore, distinct seasonal patterns are evident in both temperature and wind speed across all three cities, with peak values typically occurred during the summer months and lows during winter (**Fig. 4**). These trends align with the expected influence of solar radiation on the climatic characteristic of the northeastern United States. Wind speed patterns, particularly in New York City and Washington, D.C., also demonstrated seasonality, likely driven by regional storm dynamics. In contrast, relative humidity and barometric pressure show more moderate fluctuations. Relative humidity appeared to follow temperature trends to some extent but with more contained fluctuations, while barometric pressure remained relatively stable over time, reflecting the more conservation nature of atmospheric shifts. Wind direction showed the highest degree of variability, fluctuating considerably throughout the observation period and across locations, likely reflecting localized conditions. Overall, it highlights both the repetitive nature and the intercity differences in atmospheric conditions across the northeastern United States.

4.3 $PM_{2.5}$ concentrations

$PM_{2.5}$ values were collected daily for each air quality monitoring station, McMillian in Washington, D.C., Dudley Roxbury in Boston and Queens in New York City for 10 years (**Fig. 5**). In Washington, D.C., a maximum value of 15.68 mg/m^3 on 07/21/2021 and a minimum value of 4.06 mg/m^3 on 04/03/2020 were recorded. Boston had a maximum value of 13.94 mg/m^3 on 07/20/2011 and a minimum value of 3.34 mg/m^3 on 05/13/2020 and New York

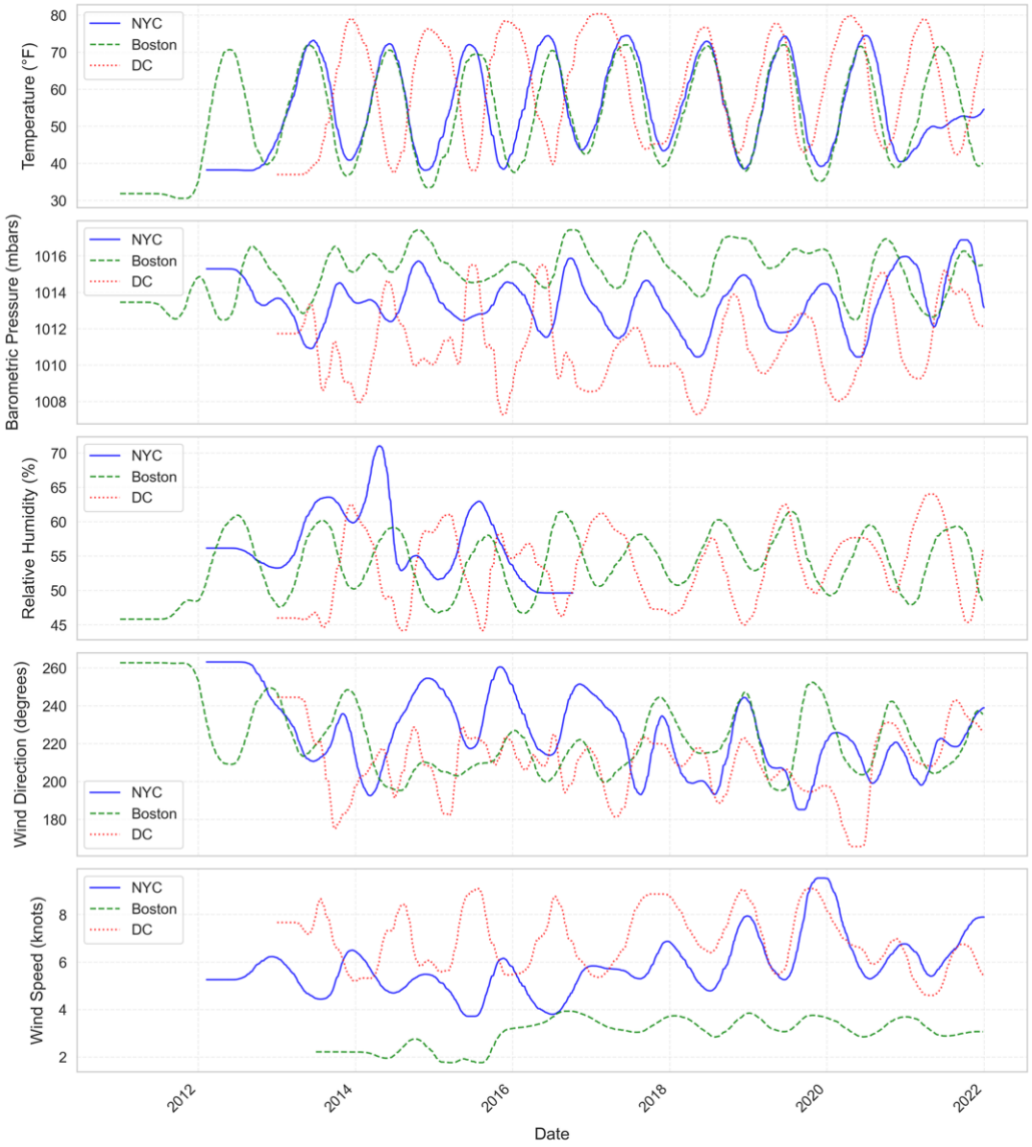


Figure 4. *Time series meteorological data for Washington, D.C., Boston and New York City.*

Table 2. *Maximum and minimum values of meteorological data throughout the 10-year study period.*

Variable	City	Maximum (date)	Minimum (date)
Temperature (°F)	Washington, D.C.	80.59 (12/15/2016)	35.90 (01/02/2013)
	Boston	72.76 (03/20/2017)	30.62 (08/29/2011)
	New York City	74.85 (04/02/2017)	37.68 (09/19/2014)
Humidity (%)	Washington, D.C.	64.13 (03/10/2021)	43.87 (07/20/2015)
	Boston	61.53 (05/16/2019)	44.88 (01/03/2011)
	New York City	72.31 (02/02/2014)	48.28 (06/10/2016)
Pressure (mbar)	Washington, D.C.	1,015 (05/27/2015)	1,007 (10/19/2015)
	Boston	1,017 (07/30/2016)	1,012 (02/21/2012)
	New York City	1,016 (07/21/2021)	1,010 (03/04/2020)
Wind Speed (knots)	Washington, D.C.	9.13 (08/25/2019)	4.58 (03/03/2021)
	Boston	3.95 (07/17/2016)	1.76 (02/14/2014)
	New York City	9.60 (09/24/2019)	3.68 (04/28/2015)
Wind Direction (°)	Washington, D.C.	247 (01/02/2013)	166 (01/02/2013)
	Boston	264 (01/03/2011)	195 (03/16/2019)
	New York City	265 (02/09/2012)	185 (07/12/2019)

City had a maximum value of 10.87 mg/m^3 on 01/06/2011 and a minimum value of 3.09 mg/m^3 on 05/09/2020. The highest $PM_{2.5}$ value in New York City suggests a major pollution event. Similarly, Boston and New York City had both experienced a peak $PM_{2.5}$ value in 2021, indicating a region-wide pollution event. All three cities experienced low $PM_{2.5}$ concentrations in 2020, suggesting a cleaner atmospheric conditions.

The occurrence of maximum $PM_{2.5}$ levels in summer months (June-July) suggests that these pollution peaks may coexist with variables like higher temperatures or increased traffic. Overall, there was an evident variability in the amount of $PM_{2.5}$ data collected across the three cities. All three cities show a significant decrease in concentrations, especially from 2020 until experiencing a drastic increase in 2021 over the 10-year period. The downward trend reflects the impact

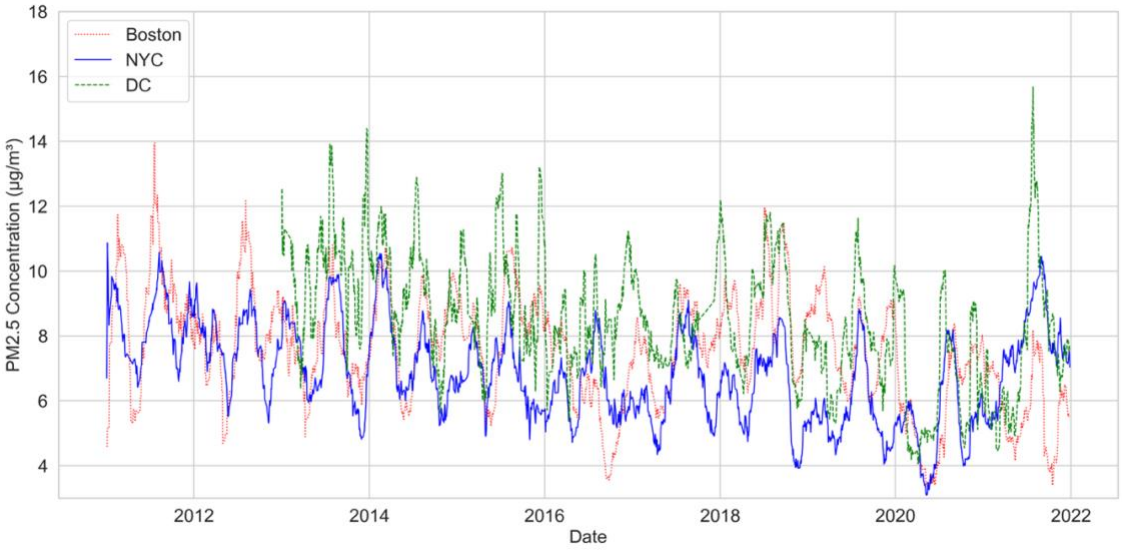


Figure 5. *Time series $PM_{2.5}$ concentrations for 10 years. Data was smoothed using a 30-day moving average to make increase readability.*

of air quality regulations, COVID pandemic, and efforts to reduce emissions from industries (Zheng et al., 2025). In the early 2020s, Boston showed several peaks, symbolizing periods of poor air quality. Similarly, Washington, D.C. and New York City have also experienced an increase in $PM_{2.5}$ levels mid-2020's. Out of all the three urban areas, Washington, D.C. had consistently moderate $PM_{2.5}$ levels throughout the period compared to Boston and New York City. Moreover, Boston showed pronounced $PM_{2.5}$ peaks because of more pollution events or poor weather conditions. This time series was a direct representation of the effectiveness of stricter air quality regulations and policies in the United States (Zheng et al., 2025).

4.4 AOD, meteorological data, and $PM_{2.5}$ relationships

Wind speed, wind direction, and relative humidity played a major role in shaping $PM_{2.5}$ concentrations throughout the 10-year period. $PM_{2.5}$ levels tended to be lower during periods of higher wind speed, which likely promoted greater atmospheric dispersion and reduced the accumulation of surface-level pollutants. In contrast, low wind speed conditions, common during colder months, may have contributed to stagnant air masses that trap pollutants near the surface, elevating $PM_{2.5}$ concentrations. Wind direction also showed a subtle but significant relationship with $PM_{2.5}$, as certain directional flows corresponded with localized spikes in fine particulate levels (**Fig. 6**). This directional dependency emphasizes the role of regional transport in urban air quality, where upwind sources can contribute to pollution. Relative humidity exhibited a positive relation with $PM_{2.5}$, particularly during the colder months. High humidity can facilitate the formation of secondary aerosols and allow particulates to absorb moisture and grow, exacerbating air quality issues. This relationship also suggests that meteorological conditions during the winter months may compound pollution effects beyond emissions alone (**Fig. 6**).

MODIS AOD, temperature and $PM_{2.5}$ concentrations each demonstrated distinct seasonal patterns across this 10-year study period. AOD values peaked mainly during the summer and early fall months (July and September), aligning with higher outdoor temperatures. This seasonal peak was also accompanied by elevated $PM_{2.5}$ concentrations, suggesting a temporal overlap in the presence of aerosols and fine particulate matter. Temperature followed an annual cycle, with maximum values in the summer and minimal values during the winter. $PM_{2.5}$ concentrations also showed seasonal variation, which increases during colder months, likely due to greater fossil fuel consumption for heating. Surprisingly, while AOD and $PM_{2.5}$ both showed summer peaks, winter $PM_{2.5}$ concentrations were also elevated despite lower AOD values, indicating that $PM_{2.5}$ sources and atmospheric behavior during colder months are not always reflected in satellite-derived AOD values. This inconsistency highlighted the complex and non-linear relationship between these variables. These observations suggest that $PM_{2.5}$ concentrations were influenced by a dynamic interaction of meteorological variables and seasonal human activity. The combination of low wind speed, high humidity, and low temperatures often marked periods of elevated pollution, particularly in winter, while higher temperatures, moderate wind speeds, and low humidity during summer correlated with elevated but more dispersed $PM_{2.5}$ levels.

4.5 $PM_{2.5}$ predictions

AOD measures the scattering and absorption of sunlight by aerosols in the atmosphere, which was related to the $PM_{2.5}$ concentration. AOD represents the total aerosol load in the atmospheric column and reflects $PM_{2.5}$ concentration of fine particulate matter near the surface. AOD data from MODIS provides comprehensive spatial and temporal coverage. Satellite-derived AOD complements surface level $PM_{2.5}$ measurements by filling gaps in monitoring networks.

Incorporating environmental variables, AOD, and temporal data into predictive models enhances their accuracy by capturing dynamic interactions between meteorological conditions and pollutant levels. The models trained only with AOD had R^2 values close to 0 and were

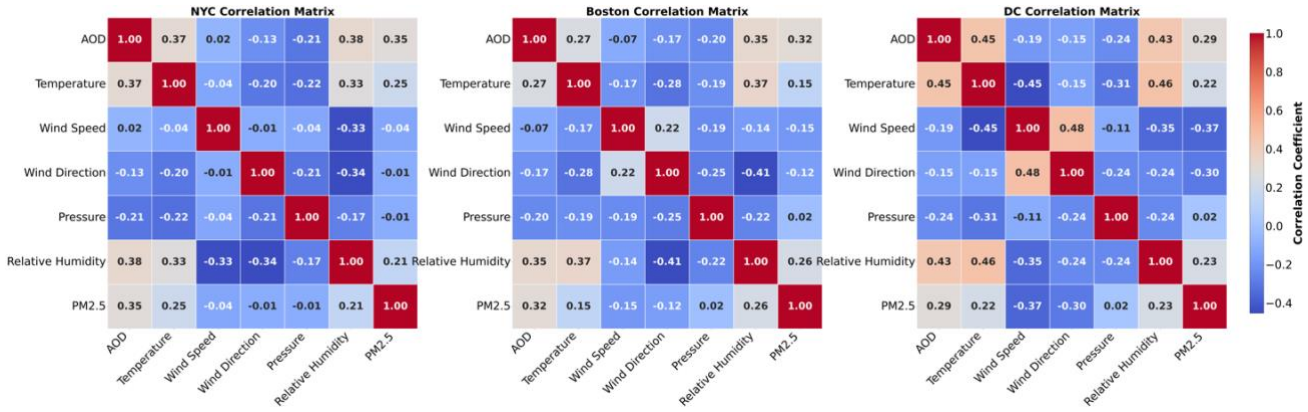


Figure 6. *Correlation matrix shows the relationship between AOD, meteorological data and PM2.5 concentrations in all three cities.*

therefore deemed poor predictors of $PM_{2.5}$ concentrations while the models trained with meteorological (wind speed, wind direction, relative humidity, temperature, and barometric pressure) and AOD data produced R^2 values of 0.25 and greater (**Tables 3–5**). The prediction was done only with models trained with meteorological and AOD

Table 3. *The results of the model training for Washington, D.C.*

Timeline	Model	R^2	MSE	MAE	RMSE
Predictors: MODIS AOD					
5 years (2017-2021)	RF	-0.06	16.07	2.87	4.01
	XGB	0.03	14.80	2.66	3.85
10 years (2011-2021)	RF	-0.04	4.99	1.63	2.23
	XGB	0.01	4.78	1.60	2.19
Predictors: MODIS AOD, meteorological data and season					
5 years (2017-2021)	RF	0.62	5.44	1.74	2.33
	XGB	0.63	5.34	1.76	2.31
10 years (2011-2021)	RF	0.52	7.32	1.89	2.71
	XGB	0.58	6.37	1.86	2.52

Table 4. *The results of the model training for Boston, Massachusetts.*

Timeline	Model	R^2	MSE	MAE	RMSE
Predictors: MODIS AOD					
5 years (2017-2021)	RF	0.07	9.87	2.37	3.14
	XGB	0.04	10.25	2.44	3.20
10 years (2011-2021)	RF	-0.06	3.86	1.47	1.97
	XGB	0.01	3.64	1.46	1.91
Predictors: MODIS AOD, meteorological data and season					
5 years (2017-2021)	RF	0.32	10.47	2.28	3.24
	XGB	0.31	10.56	2.24	3.25
10 years (2011-2021)	RF	0.30	9.29	2.15	3.05
	XGB	0.25	9.98	2.20	3.16

Table 5. *The results of the model training for Queens, New York City.*

Timeline	Model	R^2	MSE	MAE	RMSE
Predictors: MODIS AOD					
5 years (2017-2021)	RF	-0.03	16.92	3.26	4.11
	XGB	0.04	15.76	3.23	3.97
10 years (2011-2021)	RF	-0.01	2.67	1.23	1.64
	XGB	-0.01	2.65	1.24	1.63
Predictors: MODIS AOD, meteorological data and season					
5 years (2017-2021)	RF	0.39	11.55	2.44	3.40
	XGB	0.31	13.14	2.60	3.63
10 years (2011-2021)	RF	0.42	13.79	2.52	3.71
	XGB	0.34	15.69	2.67	3.96

data. To test the generalizability of the training models for each city, predictive models were developed utilizing a combination of data from two out of the three cities to predict the third city (**Table 6**). For example, the Washington, D.C. $PM_{2.5}$ predictions used New York City and Boston data, while the Boston $PM_{2.5}$ prediction model used Washington, D.C. and New York city data. The New York City $PM_{2.5}$ prediction model was trained using Boston and Washington, D.C. data.

Models using data for the 10-year period produced R^2 values of 0.68 (RF) and 0.67 (XGB) for Washington, D.C. For Boston $PM_{2.5}$ prediction models, RF and XGB models produced R^2 values of 0.21 and 0.23, respectively. For New York City $PM_{2.5}$ prediction models, RF and XGB produced R^2 values of 0.47 and 0.43, respectively (**Table 6**).

Model performance varied across locations and training periods. For Washington, D.C., both RF and XGB achieved strong results, with R^2 values improving from 0.54 (5 years) to 0.68 (10 years) for RF and from 0.54 to 0.67 for XGB, alongside decreases in error metrics (MSE, MAE and RMSE). In contrast, Boston predictions showed weaker performance for both RF, declined from 0.39 (5 years) to 0.21 (10 years), and XGB, from 0.32 to 0.23, with only minimal improvements in error values, suggesting more difficult in capturing $PM_{2.5}$ dynamics for this city. New York City results were moderate, with RF models improving slightly from 0.43 (5 years) to 0.47 (10 years), while XGB models achieved R^2 values of 0.41 to 0.43. Error metrics for New York City were higher than Washington, D.C., but lower than Boston, indicating moderate predictive ability. Overall, Washington, D.C. was the most predictable location, Boston the least and New York City fell in between (**Table 6**).

4.6 Feature importance

Feature importance was determined using both RF and XGB algorithms. Random Forest calculates feature importance based on the mean decrease in impurity, a metric that measures how much each feature reduces the impurity in the data when used to split nodes across all decision trees in the model. Each time a feature is used for a split, the resulting reduction is recorded. These reductions are then summed across all trees in the forest and normalized so that the importances across all features add up to 1 (Díaz-Uriarte and De Andres, 2006). On the other hand, XGB determines feature importance by counting the number of times a feature is used to split nodes across all boosting rounds. This count is weighted by the usefulness of the splits. As with RF, XGB importances are normalized so that their total equals 1 (Chen and Guestrin, 2016).

Feature importance indicates the key meteorological variables influencing $PM_{2.5}$ concentrations, which will provide insights into air quality dynamics across regions. The top feature demonstrates the highest importance in model predictions (Di et al., 2016). Maximum temperature, wind speed, season, mean temperature and year were the top five predictors driving model accuracy (**Figs. 7–8**). These variables influence atmospheric dispersion, chemical reactions, and pollutant accumulation. Wind speed plays a major role, as it helps in understanding how quickly

Table 6. *$PM_{2.5}$ prediction model results for Washington, D.C., Boston and New York City.*

Location	Model	Period	R^2	MSE	MAE	RMSE
Washington, D.C.	RF	5 years	0.54	9.48	2.03	3.08
	RF	10 years	0.68	6.60	1.89	2.57
	XGB	5 years	0.54	9.49	2.02	3.08
	XGB	10 years	0.67	6.69	1.88	2.59
Boston	RF	5 years	0.39	9.21	2.37	3.25
	RF	10 years	0.21	10.22	2.35	3.20
	XGB	5 years	0.32	10.22	2.32	3.20
	XGB	10 years	0.23	10.05	2.31	3.17
New York City	RF	5 years	0.43	10.55	2.20	3.25
	RF	10 years	0.47	14.81	2.87	3.85
	XGB	5 years	0.41	11.00	2.48	3.32
	XGB	10 years	0.43	13.26	2.54	3.64

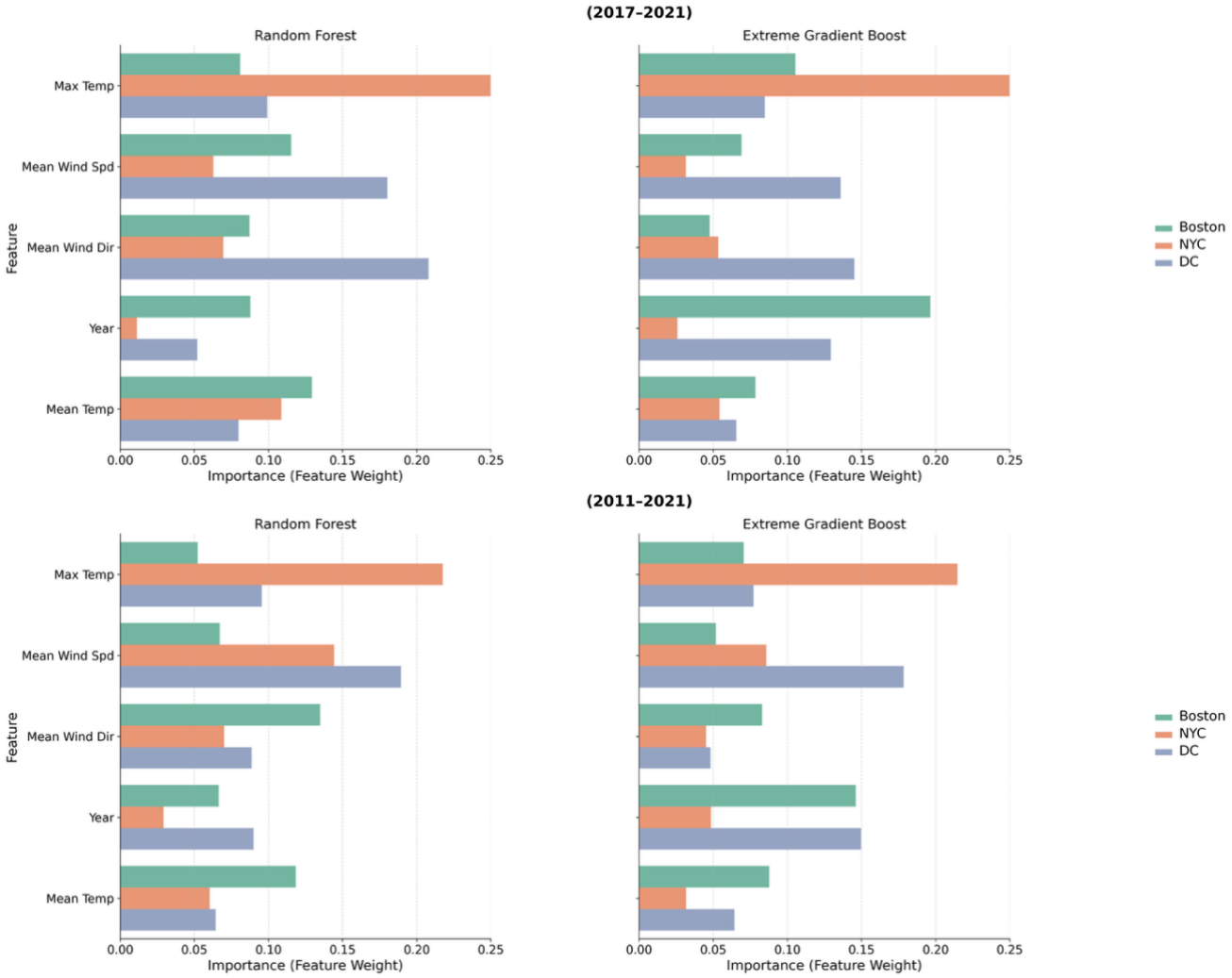


Figure 7. *Feature importances for each city across both models for the 5- and 10-year time periods.*

pollutants disperse and serves as a means of transporting pollutants across regions. Additionally, maximum temperature and mean temperature impacts human activity, e.g., during the summertime, people run air conditioning and coolers, which increases $PM_{2.5}$ concentrations (Jacob and Winner, 2009). These factors improve model performance by considering environmental variability that correlates with $PM_{2.5}$ concentrations (Zhang and Ma, 2012).

The variables 'maximum temperature', 'mean wind speed', 'year', 'mean wind direction' and 'mean temperature' emerged as important predictors of $PM_{2.5}$ concentrations, as they reflect broader climatic and temporal trends that influence air quality. Although 'year' captures long-term trends, it may not directly affect $PM_{2.5}$ concentrations as strongly as other environmental variables. In contrast, features such as maximum relative humidity, pressure, and wind direction showed lower predictive power (**Fig. 8**).

Air pollution poses a major risk to public health and environmental sustainability (Cohen et al., 2015). Accurate modeling of $PM_{2.5}$ concentrations was significant for developing effective mitigation strategies. Several major insights arose regarding the model's ability to predict $PM_{2.5}$ concentrations across different cities and periods. For all cities, the RF and XGB models built with only AOD data performed poorly across all periods (R^2 values ranging from -0.05 to 0.07 for RF and 0.01 to 0.02 for XGB). This indicates that AOD alone was a weak predictor for $PM_{2.5}$ concentrations, highlighting the need for additional variables to increase model accuracy. Incorporating meteorological data improved predictions across all periods. The highest R^2 values were observed for Washington, D.C. model results for the 2011–2021 period, with 0.68 for RF and 0.67 for XGB. The lowest R^2 values were observed in the Boston model results: 0.21 (RF) and 0.23 (XGB). Increasing data volume and variability increased the R^2 values for all models, except for Boston (**Fig. 9**). Even though, these improvements demonstrate the significant role of meteorological and MODIS AOD data in accurately predicting $PM_{2.5}$ concentrations. Boston's models seem to be an anomaly, perhaps because the variability in the city's data was too great for the generalized model to capture the underlying pattern. On the other hand, the AOD-only models for all sites have failed to produce robust predictions,

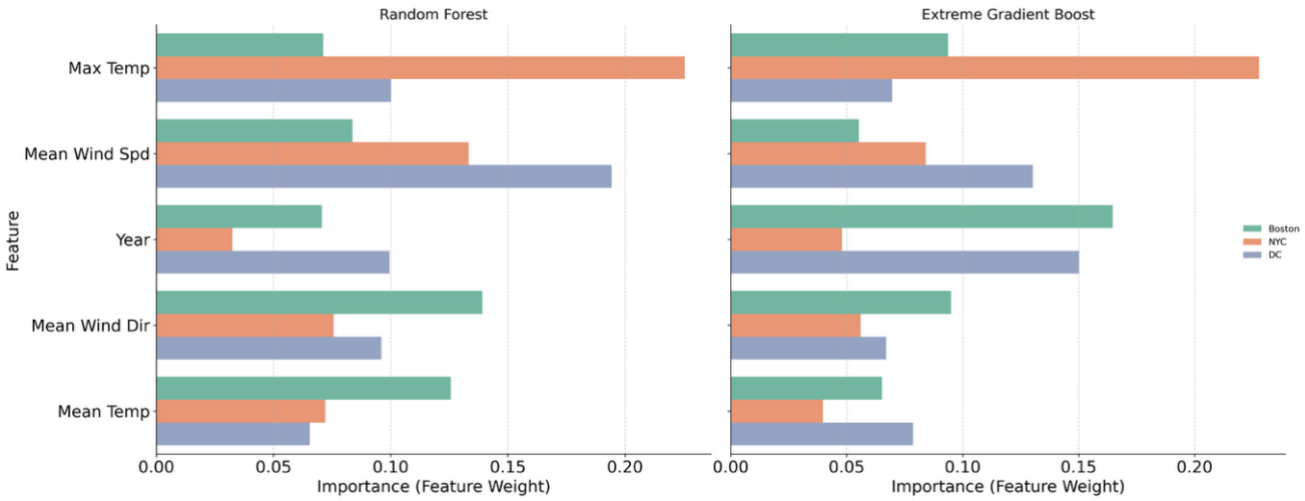


Figure 8. *Feature importances for model predictions in each city.*

even with datasets for the longer time periods (2011–2021). On the contrary, including meteorological data in the models yield better R^2 values, ultimately enhancing the predictive power. The performance of the models improved consistently as more historical data and additional variables were included into the model structure.

5 Discussion

Seasonal variability plays an important role in aerosol patterns. Similar to previous studies, we observed co-variation during summer, when high AOD, temperature, and $PM_{2.5}$ concentrations align, consistent with the effects of photochemical activity and atmospheric instability (van Donkelaar et al., 2010). In contrast, weaker correlations were observed in winter, when $PM_{2.5}$ often shows high values despite lower AOD. This creates inconsistencies in satellite-derived AOD observations under conditions such as temperature inversions and low vertical mixing. Furthermore, the relationship between temperature and $PM_{2.5}$ is complex: higher temperatures may enhance secondary aerosol formation through photochemical processes, yet they can also promote dispersion depending on the local meteorological conditions (Jacob and Winner, 2009). These findings emphasize the importance of considering seasonal and meteorological factors when interpreting urban aerosol dynamics.

Several peaks exceeding 0.30 AOD were recorded in New York City and Washington, D.C. during 2020 and 2021, indicating periodic episodes of elevated aerosol concentrations. These spikes are commonly associated with environmental events such as dust storms or wildfires, which have been shown to significantly increase AOD levels in urban areas (Daniels et al., 2024). Such patterns highlight the need for multi-variable approaches when modeling air quality in urban areas.

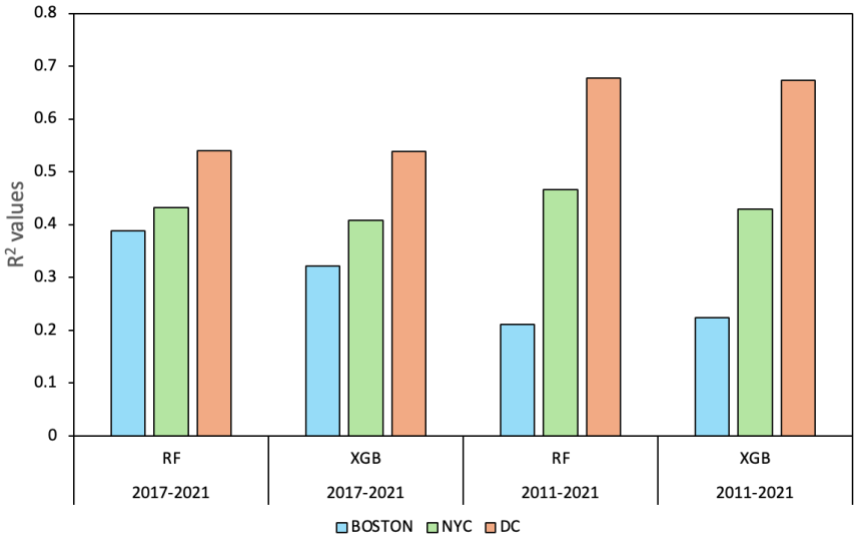


Figure 9. *$PM_{2.5}$ predictive model performances across study areas and model types.*

The use of AOD measurements at a 550 nm wavelength as a predictor is well established in the literature. This wavelength shows a strong correlation with $PM_{2.5}$ concentrations in urban areas and is minimally influenced by non-aerosol atmospheric components, making it a reliable choice for remote sensing applications (Handschuh et al., 2022). However, previous studies have emphasized that AOD alone is insufficient for accurate $PM_{2.5}$ prediction due to spatial heterogeneity, vertical aerosol profiles, and meteorological variability (Gupta et

al., 2006). These results highlighted the need for multi-parameter approaches in air quality modelling.

Our model results highlighted the complex nature of aerosol-meteorological processes. The relatively low R^2 values reflect the inherent challenges of predicting surface-level $PM_{2.5}$ from satellite-derived AOD and meteorological data. This outcome is consistent with previous studies. For instance, Liu et al. (2007) reported R^2 values of 0.36–0.55 for MODIS AOD and $PM_{2.5}$ relationships in the northeastern United States, while Paciorek and Liu (2009) observed R^2 values as low as 0.2–0.4 across multiple regions. Similarly, van Donkelaar et al. (2010) noted persistent challenges in urban-scale prediction. Zheng et al. (2021) and Kumar and Pande (2023) observed that their machine learning models produced moderate R^2 values (0.3–0.6) depending on season and geography. Within this context, our results are comparable, and importantly, it showed a clear improvement relative to AOD-only models. The addition of meteorological variables improved explained variance and reduced RMSE by 20–30%, demonstrating meaningful predictive power even if absolute R^2 values remain modest. These findings emphasized both the utility and limitations of AOD-based modeling and highlight the importance of continued methodological development. By building on the framework presented here with higher-resolution data, chemical transport model outputs, and deep learning approaches, future research can more effectively support air quality monitoring and policy planning in data-sparse and vulnerable regions.

Several research directions can be considered: firstly, extending predictions to areas without monitoring stations is critical for addressing gaps in air quality data coverage. This could involve combining remote sensing, meteorological inputs, and spatial interpolation techniques with geostatistical and machine learning methods. Incorporating land-use information may further improve model accuracy by accounting for localized emission sources; secondly, integrating $PM_{2.5}$ prediction models with climate projections would provide insight into how changing temperature, wind patterns, and precipitation regimes may affect future air quality. Such projections are especially valuable for informing long-term mitigation strategies. Given that marginalized and low-income communities often experience the greatest burden of poor air quality, these efforts carry important implications for reducing health disparities and addressing socio-political inequities in urban environments (Tessum et al., 2019; Josey et al., 2023).

6 Conclusion

This study demonstrated that AOD data alone provided weak capabilities for $PM_{2.5}$ predictions in urban areas like Washington, D.C., Boston, and New York City. Although the integration of MODIS AOD and meteorological data significantly improved all model performance, there were limitations in data availability. The results highlight the importance of integrating satellite-based observations (AOD) with ground-based meteorological measurements to improve the accuracy of $PM_{2.5}$ predictions. In addition, the findings reveal a weak correlation between AOD and $PM_{2.5}$ concentrations and demonstrate the performance of RF and XGB models across multiple years. The models built using both dataset types have proven to be dependable. The predictive accuracy improved with datasets for longer time periods (2011-2021), particularly in Washington, D.C., suggest that these models can provide valuable insights for air quality management in major urban environments. Furthermore, model development could explore additional meteorological variables or possibly alternative remote sensing data to further improve prediction accuracy.

7 Acknowledgement

First author would like to thank Professor Sun for his guidance during graduate thesis works. I would also like to thank Jalen Grandchamps for his unwavering support during writing this manuscript.

8 Data availability statement

The data that supports this research will be shared upon reasonable request to the corresponding authors.

9 Author contributions

JAM: conceptualization, data curation, formal analysis, visualization, and writing – original draft. WNM: methodology, resources, supervision, and writing – review & editing. MR: methodology, validation, and writing – review & editing. All authors approved the final version of the manuscript.

10 Conflict of interest

The authors declare that they have no known competing financial Interests or personal relationships that could have appeared to influence the work reported in this paper.

11 Ethical statements

Not applicable.

12 Copyright statement

This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY NC ND) license (https://creativecommons.org/licenses/by-nc-nd/4.0/). © 2025 by the authors. Licensee Enviro Mind Solutions, Connecticut, USA.

References

American Lung Association, 2024. State of the air: District of Columbia. https://www.lung.org/research/sota/city-rankings/states/district-of-columbia/district-of-columbia (Accessed on August 29, 2024).

Bărbulescu, A., Dumitriu, C.S., Ilie, I., Barbeş, S.-B., 2022. Influence of anomalies on the models for nitrogen oxides and ozone series. Atmosphere, *13*, 558. https://doi.org/10.3390/atmos13040558

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

Chu, Y., Liu, Y., Li, X., Liu, Z., Lu, H., Lu, Y., Mao, Z., Chen, X., Li, N., Ren, M., Liu, F., Tian, L., Zhu, Z., Xiang, H., 2016. A review on predicting ground $PM_{2.5}$ concentration using satellite aerosol optical depth. Atmosphere, 7, 129. https://doi.org/10.3390/atmos7100129

Cohen, A.J., Ross Anderson, H., Ostro, B., Pandey, K.D., Krzyzanowski, M., Künzli, N., Gutschmidt, K., Pope, A., Romieu, I., Samet, J.M., Smith, K., 2005. The global burden of disease due to outdoor air pollution. Journal of Toxicology and Environmental Health, Part A, 68, 1301–1307. https://doi.org/10.1080/15287390590936166

Daniels, J., Liang, L., Benedict, K.B., Brahney, J., Rangel, R., Weathers, K.C., Ponette-González, A.G., 2024. Satellite-based aerosol optical depth estimates over the continental U.S. during the 2020 wildfire season: Roles of smoke and land cover. Science of The Total Environment, 921, 171122. https://doi.org/10.1016/j.scitotenv.2024.171122

Díaz-Uriarte, R., Alvarez de Andres, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7, 3. https://doi.org/10.1186/1471-2105-7-3

Di, Q., Koutrakis, P., Lyapustin, A., Wang, Y., Schwartz, J., 2016. Assessing $PM_{2.5}$ exposures with high spatiotemporal resolution across the continental United States. Environmental Science & Technology, 50, 4712–4721. https://doi.org/10.1021/acs.est.5b06121

Feng, S., Gao, D., Liao, F., Zhou, F., Wang, X., 2016. The health effects of ambient $PM_{2.5}$ and potential mechanisms. Ecotoxicology and Environmental Safety. 128, 67–74. https://doi.org/10.1016/j.ecoenv.2016.01.030

Gupta, P., Christopher, S.A., 2008. Seven year particulate matter air quality assessment from surface and satellite measurements. *Atmospheric Chemistry and Physics*, 8, 3311–3324. https://doi.org/10.5194/acp-8-3311-2008

Gupta, P., Christopher, S.A., Wang, J., Gehrig, R., Lee, Y., Kumar, N., 2006. Satellite remote sensing of particulate matter and air quality assessment over global cities. *Atmospheric Environment*, 40, 5880–5892. https://doi.org/10.1016/j.atmosenv.2006.03.016

Gutiérrez-Avila, I., Arfer, K.B., Carrión, D., Rush, J., Kloog, I., Naeger, A.R., Grutter, M., Páramo-Figueroa, V.H., Riojas-Rodríguez, H., Just, A.C., 2022. Prediction of daily mean and one-hour maximum $PM_{2.5}$ concentrations and applications in Central Mexico using satellite-based machine-learning models. Journal of Exposure Science & Environmental Epidemiology, 32, 917–925. https://doi.org/10.1038/s41370-022-00471-4

Handschuh, J., Erbertseder, T., Schaap, M., Baier, F., 2022. Estimating $PM_{2.5}$ surface concentrations from AOD: A combination of SLSTR and MODIS. Remote Sensing Applications: Society and Environment, 26, 100716. https://doi.org/10.1016/j.rsase.2022.100716

Jacob, D.J., Winner, D.A., 2009. Effect of climate change on air quality. *Atmospheric Environment*, 43, 51–63. https://doi.org/10.1016/j.atmosenv.2008.09.051

Jaffe, D.A., O'Neill, S.M., Larkin, N.K., Holder, A.L., Peterson, D.L., Halofsky, J.E., Rappold, A.G., 2020. Wildfire and prescribed burning impacts on air quality in the United States. *Journal of the Air & Waste Management Association*, 70, 583–615. https://doi.org/10.1080/10962247.2020.1749731

Josey, K.P., Delaney, S.W., Wu, X., Nethery, R.C., DeSouza, P., Braun, D., Dominici, F., 2023. Air pollution and mortality at the intersection of race and social class. The New England Journal of Medicine, 388, 1396–1404. https://doi.org/10.1056/nejmsa2300523

Karner, A.A., Eisinger, D.S., Niemeier, D.A., 2010. Near-roadway air quality: Synthesizing the findings from real-world data. Environmental Science & Technology, 44, 5334–5344. https://doi.org/10.1021/es100008x

Kaveh, M., Mesgari, M.S., Kaveh, M.A., 2025. Novel evolutionary deep learning approach for $PM_{2.5}$ prediction using remote sensing and spatial–temporal data: A case study of Tehran. International Journal of Geo-Information, 14, 42. https://doi.org/10.3390/ijgi14020042

Kibirige, G.W., Yang, M.C., Liu, C.L., Chen, M.C., 2023. Using satellite data on remote transportation of air pollutants for $PM_{2.5}$ prediction in northern Taiwan. PLOS ONE, 18, e0282471. https://doi.org/10.1371/journal.pone.0282471

Kumar, K., Pande, B.P., 2023. Air pollution prediction with machine learning: A case study of Indian cities. International Journal of Environmental Science and Technology, 20, 5333–5348. https://doi.org/10.1007/s13762-022-04241-5

Kumar, N., Chu, A., Foster, A., 2007. An empirical relationship between $PM_{2.5}$ and aerosol optical depth in Delhi Metropolitan. Atmospheric Environment, 41, 4492–4503. https://doi.org/10.1016/j.atmosenv.2007.01.046

Li, J, An, X., Li, Q., Wang, C., Yu, H., Zhou, X., Geng, Y.A., 2022. Application of XGBoost algorithm in the optimization of pollutant concentration. *Atmospheric Research*, 276, 106238. https://doi.org/10.1016/j.atmosres.2022.106238

Liu, Y., Franklin, M., Kahn, R., Koutrakis, P., 2007. Using aerosol optical thickness to predict ground-level $PM_{2.5}$ concentrations in the St. Louis area: A comparison between MISR and MODIS. Remote Sensing of Environment, 107, 33-44. https://doi.org/10.1016/j.rse.2006.05.022

Nath, B., Chowdhury, R., Ni-Meister, W., Mahanta, C., 2022. Predicting the distribution of arsenic in groundwater by a geospatial machine learning technique in the two most affected districts of Assam, India: The public health implications. GeoHealth, 6, e2021GH000585. https://doi.org/10.1029/2021GH000585

Paciorek, C.J., Liu, Y., 2009. Limitations of remotely sensed aerosol as a spatial proxy for fine particulate matter. Environmental Health Perspectives, 117, 904-909. https://doi.org/10.1289/ehp.0800360

Park, Y., Kwon, B., Heo, J., Hu, X., Liu, Y., Moon, T., 2020. Estimating PM2.5 concentration of the conterminous United States via interpretable convolutional neural networks. Environmental Pollution, 256, 113395. https://doi.org/10.1016/j.envpol.2019.113395

Qin, Y., Kim, E., Hopke, P.K., 2006. The concentrations and sources of $PM_{2.5}$ in metropolitan New York City. Atmospheric Environment, 40, 312–332. https://doi.org/10.1016/j.atmosenv.2006.02.025

Remer, L.A., Kaufman, Y.J., Tanré, D., Mattoo, S., Chu, D.A., Martins, J.V., Li, R., Ichoku, C., Levy, R.C., Kleidman, R.G., Eck, T.F., Vermote, E., Holben, B.N., 2005. The MODIS aerosol algorithm, products, and validation. Journal of the Atmospheric Sciences, 62, 947–973. https://doi.org/10.1175/JAS3385.1

Samad, A., Garuda, S., Vogt, U., Yang. B., 2023. Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations. *Atmospheric Environment*, 310, 119987. https://doi.org/10.1016/j.atmosenv.2023.119987

Tessum, C.W., Apte, J.S., Goodkind, A.L., Muller, N.Z., Mullins, K.A., Paolella, D.A., Polasky, S., Springer, N.P., Thakrar, S.K., Marshall, J.D., Hill, J.D., 2019. Inequity in consumption of goods and services adds to racial–ethnic disparities in air pollution exposure. *Proceedings of the National Academy of Sciences U.S.A.*, 116, 6001–6006. https://doi.org/10.1073/pnas.1818859116

van Donkelaar, A., Martin, R.V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., Villeneuve, P.J., 2010. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth. Environmental Health Perspectives, 118, 847–855. https://doi.org/10.1289/ehp.0901623

Wong, P.Y., Su, H.J., Lee, H.Y., Chen, Y.C., Hsiao, Y.P., Huang, J.W., Teo, T.A., Wu, C.D., Spengler, J.D., 2021. Using land-use machine learning models to estimate daily NO_2 concentration variations in Taiwan. Journal of Cleaner Production, 317, 128411. https://doi.org/10.1016/j.jclepro.2021.128411

Zhang, C., Ma, Y., 2012. Ensemble Machine Learning: Methods and Applications. Springer Publishing Company. https://doi.org/10.1007/978-1-4419-9326-7

Zheng, M., Liu, F., Wang, M., 2025. Assessing the COVID-19 lockdown impact on global air quality: A transportation perspective. Atmosphere, 16, 113. https://doi.org/10.3390/atmos16010113

Zheng, T., Bergin, M., Wang, G., Carlson, D., 2021. Local $PM_{2.5}$ hotspot detector at 300 m resolution: A random forest–convolutional neural network joint model jointly trained on satellite images and meteorology. Remote Sensing, 13, 1356. https://doi.org/10.3390/rs13071356

Publisher's note

The author(s) are solely responsible for the opinions and data presented in this article, and publisher or the editor(s) disclaim responsibility for any injury to people or property caused by any ideas mentioned in this article.